2026, Volume 58, Issue 2, 81-87, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Overview of Quantitative Research

TingLan Ma, PhD; Yen Lee, PhD

AUTHOR AFFILIATION:
Department of Health Professions Education, Uniformed Services University, Bethesda, MD

CORRESPONDING AUTHOR:
TingLan Ma, Department of Health Professions Education, Uniformed Services University, Bethesda, MD, ting-lan.ma.ctr@usuhs.edu

HOW TO CITE: Ma T, Lee Y. Overview of Quantitative Research. *Fam Med.* 2026;58(2):81-87.
doi: 10.22454/FamMed.2026.406133

FIRST PUBLISHED: February 12, 2026

KEYWORDS: correlational, experiment, observation, quantitative research, study design, survey

ABSTRACT

Quantitative research helps medical educators and researchers use data to understand and improve learning, teaching, and program outcomes. Applying statistical methods to summarize and compare results makes it possible to measure change, identify patterns, and evaluate educational efforts, such as new curricula, wellness initiatives, or assessment of programs. This article introduces key ideas for using quantitative methods effectively in medical and family medicine education, including how research questions connect to study design, common approaches such as experimental, quasi-experimental, and correlational studies, and practical ways to collect data through surveys, observations, or existing records. Examples from medical education illustrate how these methods can be used to evaluate programs, describe learner progress, and test innovations. The paper also outlines common challenges—such as drawing broad conclusions from small samples, confusing association with cause, or using measures that do not fully capture what is intended—and offers strategies to address these problems. The paper aims to help clinician-educators apply quantitative methods with greater confidence and clarity.

Quantitative research is a structured, empirically grounded approach that goes beyond simply gathering numerical data; it employs theoretically informed methodologies to test hypotheses and investigate relationships, patterns, and trends. Quantitative research offers various approaches one can choose to answer the research questions and understand the data, yet each approach comes with distinct advantages and limitations. This overview provides a general introduction to quantitative research and various approaches to a quantitative study, as well as discusses common pitfalls and strategies to address them.

IMPORTANT CONSIDERATIONS ABOUT QUANTITATIVE RESEARCH

Quantitative methods are used when researchers aim to measure variables, test hypotheses, or evaluate relationships and outcomes—particularly when the goal is to determine *how much*, *how often*, or *to what extent* something occurs. Quantitative approaches are appropriate when researchers seek to generate measurable evidence about learning processes, performance, or program effectiveness. In practice, quantitative studies often address three types of questions: (a) descriptive, which summarize collected data in meaningful ways (eg, reporting average patterns from program evaluation data or quality improvement project); (b) causal, which estimate treatment or an intervention's effects (eg, examining the impact of a new curriculum on learner outcomes); and (c) associative, which describe relationships between variables (eg, exploring how clinical competency is associated with trainees' professional identity or values of professionalism).

Once the type of question is established, researchers must consider the intended scope of their findings—whether the goal is to improve a local program or to produce results that can inform broader educational contexts. Different

choices of study design determine how results are interpreted. For example, a program may conduct a small evaluation study to inform curricular changes within its own setting. Such descriptive studies are valuable for identifying local trends and guiding program improvement, even though their findings are not intended to be generalized beyond that context. In one study, researchers evaluated family-oriented (FO) attitudes and observed skills of family medicine residents before and after a 20 week psychosocial medicine curriculum.[1] They found that stronger FO attitudes and modest increases in related behaviors were identified following the curriculum, which can be used to guide local curricular refinements.[1] Similarly, another study analyzed survey data from graduates of a single family medicine residency program to examine factors associated with focused practice.[2] Their findings—linking postgraduate year three training completion to adopting a focused practice approach and ranges of services provided—offered important insights to local institution as they refined the training policy.

In other cases, researchers pursued studies aimed at generalizability—that is, producing findings that can reasonably be applied to populations or settings beyond the original study context. Generalizable studies typically require sampling strategies and designs that reduce the chance that findings apply only to one program and instead reflect patterns likely to hold true in other settings—a key aspect of external validity.[3] This type of research often includes multisite quantitative studies that test whether similar curricular interventions yield consistent outcomes across settings, or instrument-development studies designed to broadly evaluate learner outcomes. For example, Dyrbye and colleagues developed a well-being index using data from 2248 medical students across seven institutions.[4] Careful attention to sampling and psychometric testing ensured that the index could validly measure well-being across diverse medical school contexts and future learner populations.

DIFFERENT APPROACHES OF STUDY DESIGNS

Within quantitative research, researchers may consider using different approaches of study designs to answer research questions.

Experimental Designs

Experimental studies are used when researchers want to determine whether a specific intervention (eg, a curriculum) directly causes a change in outcomes. The most rigorous form of experimental design is the randomized controlled trial (RCT), in which participants—or groups such as residency cohorts or clinic sites—are randomly assigned to different conditions to balance preintervention differences between groups. Studies have applied RCTs to evaluate intervention strategies targeting physician burnout across specialties.[5] In such a design, one group may participate in a new curricular or wellness program (the intervention condition), while another group follows standard training (the comparison condition). A strong experimental study is characterized by internal validity, meaning that any observed changes in outcomes can be attributed to intervention completion (ie, the independent variable) rather than to other confounding factors.[6] This approach requires establishing temporal precedence, with outcomes measured after the intervention to confirm causality. By carefully structuring when, how, and in what form interventions occur, researchers can minimize confounding variables and reduce alternative explanations.

The success of an experimental design also depends on maintaining control over the intervention's core elements, including consistency of content, delivery, timing, and structure. Because of this high level of control, experimental studies often can yield reliable conclusions even with modest sample sizes (eg, 30 participants per condition). However, the extent to which findings can be applied to other training programs—known as external validity—depends on how representative the sample and study conditions are. While experimental designs are powerful for establishing cause-and-effect relationships, their highly controlled nature may limit generalizability to family medicine learning or practice environments.

Quasi-Experimental Designs

Researchers often work in complex learning and clinical environments where full experimental control is not possible. When researchers can introduce an intervention, such as a new curriculum, but cannot fully control variables in the setting (eg, year-to-year variations in student composition and academic performance, or when clinical sites differ in patient population, teaching approach, or faculty support), a quasi-experimental design may be appropriate. These designs share the logic of experimentation but differ from true experiments in the degree of control researchers can exert over the learning environment, participant assignment, or contextual factors. Randomization may even be included in some quasi-experiments, such as when intact groups (eg, residency cohorts, clinic teams, or clerkship sites) are randomly assigned to receive different educational interventions. However, researchers still lack full control over participant-level exposure and contextual variation across settings, which keeps these designs distinct from true experiments. In other situations, randomization is not feasible, and groups are determined by practical or institutional constraints (eg, residents assigned to different clinical sites).

The strength of such a design depends on how well confounding factors are measured and addressed. Carefully determining key measures, such as residents' initial clinical reasoning scores or prior exposure to similar curricula, is essential to control for group differences and to minimize alternative explanations for the results. For example, a family medicine residency might implement a new peer-coaching model in one cohort and compare outcomes (eg, clinical reasoning scores or well-being) with another cohort not

receiving the intervention, while statistically adjusting for prior performance, training experience, and burnout levels to rule out their influence.

Correlational (Observational) Designs

At times, researchers are not attempting to establish causation but to describe how variables are associated in naturally occurring data, without researcher manipulation of exposures or conditions. For example, one might examine whether participation in mentorship programs is related to self-efficacy,[7] whether residents' sense of belonging predicts burnout levels, or whether faculty feedback frequency is associated with residents' clinical confidence. Because no variables are manipulated, correlational studies are nonexperimental, and their estimates describe associations rather than causal effects.[8-10] Correlational designs are relatively easy to implement, practical, and accessible, making them popular among researchers. These studies are particularly useful in applied educational settings, where randomization and having control are often impractical but understanding relationships among training, attitudes, and outcomes can inform program improvement. This design requires careful model specification and rigorous statistical control (eg, including and controlling for demographic variables) to account for potential confounding factors and measurement biases.

DATA COLLECTION METHODS

A key consideration is *how* data will be collected—whether through self-report, observation, standardized assessments, or existing data sources. This consideration can be viewed from three perspectives: whether the data are self-reported, whether data collection involves interaction with participants, and whether data are gathered at a single point in time or repeatedly over a period. Cross-sectional studies collect data from a sample at a single point in time, providing a snapshot of existing relationships or characteristics. In contrast, longitudinal studies use repeated measurements to track changes within the same participants over time, offering insights into developmental or temporal trends—for example, how in-training examination scores in residency change over time.

Primary Versus Secondary Data Sources

The major difference between primary data and secondary data analysis is whether new data are being collected. *Primary data collection* involves gathering original data through methods such as surveys, experiments, or structured observations. *Secondary data analysis*, by contrast, uses preexisting datasets that have already been collected for another purpose. This latter approach is increasingly common in medical education research, where researchers leverage curated datasets—such as the Council of Academic Family Medicine Educational Research Alliance survey database, Association of American Medical Colleges graduation questionnaire, or institutional learner assessment repositories—to examine trends, test new hypotheses, or explore predictors of outcomes without collecting new data. Secondary analyses can provide high-value insights at lower cost and with faster turnaround while still allowing for robust quantitative inquiry.

Observation and Survey as Data Collection Methods

Observation refers to a data-collection method in which researchers systematically record behaviors, interactions, or environmental features as they occur in natural or structured settings. Observational data collection relies on naturally occurring data (no researcher-participant interaction) to explore trends and patterns; for example, researchers have used video-based observation in primary care to analyze clinician-patient communication patterns.[11] Structured observation protocols and reliability checks (eg, interrater reliability)[12] help ensure consistency in how behaviors are categorized and interpreted.

Surveys are among the most widely used quantitative data-collection tools, in which structured questions are designed to capture participants' self-reported attitudes, perceptions, or experiences through standardized response options or rating scales.[13] One study illustrated this approach by surveying medical students about their confidence in addressing social determinants of health and their preparedness for advocacy.[14] Recommendations also outline best practices for survey design and implementation.[13] A systematic process—such as conducting a literature review, using interviews or focus groups to capture the target population's language, obtaining expert validation on item clarity and relevance, and performing cognitive interviews to ensure intended interpretation—can strengthen both measurement quality and help reduce common errors in survey design (eg, ambiguous wording or inconsistent response scales).

As data-collection methods, both observational and survey methods can be implemented in experimental, quasi-experimental, or correlational research designs. Meanwhile, some researchers may use the terms *observational study* and *survey study* to describe a stand-alone research design in which these tools are the primary source of data (eg, a longitudinal survey study design).[15]

RECOGNIZING DIFFERENT GOALS FOR QUANTITATIVE RESEARCH

Quantitative research may be theory-driven or data-driven. A confirmatory approach aims to validate existing theories, whereas the exploratory approach seeks to identify new patterns or relationships.

Confirmatory Approach

When the goal is to validate and refine theories from other fields (eg, examine the effects of a curriculum intervention), a confirmatory approach is practical. In such work, the links among theory, hypotheses, and result testing guide study design and interpretation.[16] For example, theoretical

frameworks such as self-determination theory,[17] cognitive load theory,[18] and theories of identity formation,[19,20] offer a conceptual foundation for formulating hypotheses that predict learner outcomes. Hypotheses derived from established theories are then tested to determine whether empirical evidence supports or fails to support them. For example, in a study where experiential learning theory informed the research question, researchers assessed whether different teaching modalities—such as theater in education, simulated patients, and role-play—were equally effective in enhancing communication skills among second-year medical students.[21] Another study used theoretical modeling to evaluate whether a self-directed learning curriculum improves clinical competency.[22] This top-down, confirmatory, and theoretical-driven approach helps refine curricula and determine whether a prespecified learning model holds or needs revision.

Exploratory Approach

Exploratory research identifies patterns or relationships when little prior knowledge exists. For example, in the same experiential learning study,[21] researchers also explored gender differences in communication skills following different teaching modalities without a guided theory. This exploration allowed them to evaluate whether various teaching modalities work similarly well for learners across different demographic backgrounds. Exploratory results may also generate new hypotheses. For instance, one study analyzed survey data to examine correlations between preferred learning modalities and burnout among physician assistant students;[23] results revealed new areas for targeted interventions. Such findings can extend existing theories or serve as foundations for new models to be tested in later confirmatory studies.

Often, researchers use a mixed approach that combines exploratory and confirmatory methods—drawing on established predictors while allowing data-driven insights to emerge. This blended strategy enables testing theory-driven questions while remaining open to context-specific patterns that refine subsequent hypotheses. For example, a study on medical students' coping strategies began with coping theory to guide analyses, then identified strategies unique to the medical school environment.[24]

COMMON PROBLEMS IN QUANTITATIVE RESEARCH

Reflecting on the limitations of quantitative research is important. While this research offers valuable insights, it faces challenges such as statistical misinterpretations, overgeneralization, and issues in measurement quality and social desirability bias.

Overgeneralization and Misuse of Causal Language

A common issue is *overgeneralization*, where results from a specific sample (eg, students in one institution) are applied too broadly. Small-scale studies using primary data are particularly prone to this problem, but it can be mitigated by providing detailed demographic descriptions of the sample and contextual information (both historical and geographic) about the study site. This information enables readers to assess the transferability of findings to their own settings. Authors also should be careful not to overstate the broader significance of a single study's results; conclusions should reflect the scope and scale of the evidence presented.

Overuse of causal language is another frequent issue when studies with observational or cross-sectional designs report findings in causal terms.[25,26] For example, a study might find that students who engage in active learning perform better on exams; but without an experimental design, one cannot conclude that active learning caused the improved performance. This error often arises from using terms like "impact" or "influence" to describe associations between variables. Even in longitudinal designs where surveys are administered repeatedly, variables measured earlier cannot be definitively interpreted as causes of later outcomes. Such studies are better described as longitudinal correlational, and terms like "longitudinal association" should be used. Researchers must align conclusions with study design and clearly distinguish correlation from causation.

Statistical Misinterpretations and *P* Value Myths

A common misconception is the overreliance on *P* values to determine significance. A *P* value less than .05 does not prove that a hypothesis is true; rather, it represents the probability of obtaining a result as extreme as (or more extreme than) the observed one, assuming that the null hypothesis is true. In other words, a small *P* value indicates that such a result would be unlikely to occur by chance across repeated samples if no true effect exists. Likewise, a nonsignificant *P* value does not confirm the absence of a relationship; it may simply reflect low power, often due to small sample size. Instead of focusing solely on *P* values, researchers may report confidence intervals, model fit indicators, and effect sizes for a more nuanced interpretation of results.

Additionally, conducting multiple statistical tests without adjusting the study Type I error rate increases the likelihood of false positives. Some studies have reported dozens of significance tests, which inflated the overall Type I error (the probability of incorrectly rejecting at least one true null hypothesis). To mitigate this problem, ideally researchers would report all performed tests, include effect sizes,[27] and, most importantly, apply proper alpha corrections when multiple comparisons are conducted.[28-31]

Myth About Equal Sample Size

In research, unequal sample sizes often occur when collecting survey data in real-world settings. When comparing unequal groups (eg, 100 male participants vs 200 female participants), the statistical power of the analysis is primarily affected by the size of the smaller group, and thus the

power to detect significant differences may be reduced.[32] However, this difference in sample size does not imply that group comparisons are invalid.[33] Rather, the results remain interpretable and meaningful.

Problems in Measurement Quality

In survey-based quantitative studies, a recurring challenge is ensuring measurement quality. Two core aspects are reliability—the consistency of an instrument, and construct validity—the extent to which it measures what it is intended to measure. Problems arise when instruments conflate related but distinct domains. For instance, a survey intended to assess communication skill development may capture learners' confidence or intentions rather than behavioral change, and a clinical competency scale may measure perceived proficiency rather than performance. Such mismatches highlight the need to critically evaluate whether an instrument aligns with the intended construct, rather than assuming that a previously validated tool always measures what it claims.

Even with existing measures, reliability and validity are not automatically transferable across contexts—a limitation referred to as measurement transferability. Because evidence of reliability and validity is tied to the context in which it was established, an instrument that performs well in one setting may not function the same way in another. For example, a professionalism scale validated with practicing physicians may not be equally reliable or valid for postgraduate trainees, and is even less appropriate for first-year medical students, whose developmental stage and interpretations of professionalism differ substantially. Thus, researchers should not rely solely on previously reported reliability and validity but should consider contextual differences and, when possible, evaluate instruments in their own setting.

Social Desirability and Acquiescence Biases

Response biases are common in self-report measures. Social desirability bias[34] occurs when participants provide answers they believe will be viewed favorably by others. For instance, students may overstate patient-centered attitudes to meet perceived expectations of faculty. Acquiescence bias, by contrast, reflects a general tendency to agree with statements regardless of content.[35] Although distinct, these biases can overlap, especially when positively worded items align with socially valued behaviors. For example, residents might overreport patient-centered attitudes or professional values, either to align with perceived norms (social desirability bias) or simply to agree with affirming statements (acquiescence bias). To address these issues, researchers sometimes include social desirability scales as covariates,[36,37] but their effectiveness varies.[34] Balancing positively and negatively worded items, ensuring response anonymity, and using neutral phrasing also can help reduce acquiescence.

Triangulating self-reports with complementary data sources, such as clinical performance evaluations, patient feedback, faculty assessments, or direct observations may minimize these issues.[38-40] While multisource data (eg, 360 degree evaluations) can enhance research validity,[41,42] collecting that data is not always appropriate or feasible.[34,43,44] This complexity becomes profound where evaluating clinical competence requires input from peers, faculty, and patients; perspectives that may not always align.[45] For example, a resident's self-assessment of their communication skills may differ significantly from a patient's or a faculty's perception.[45]

When Self-report Matters

While performance-based measures can reduce certain reporting bias, they may not fully capture the nuanced experiences central to educational, psychological, and affective constructs. In areas such as burnout, well-being, emotion regulation, or mistreatment, learners' interpretations and personal perspectives provide essential insight into mental health outcomes (eg, depression).[46] Researchers should carefully design measurements that align with the constructs and the study goals. Because affective and behavioral domains evolve over time, longitudinal follow-up with repeated measures can strengthen internal validity by capturing temporal change and reducing reliance on single time-point self-reports.

CONCLUSIONS

Quantitative research has advanced medical education, but its value and applicability rely on rigorous design, measurement, and analysis that align with research questions. Treating different study designs (eg, experimental, correlational) equivalently and using statistical controls to compensate for weak study design can lead researchers to claim more than their data really show. Pressure to demonstrate impact often pushes researchers to use cause-and-effect language even when data only show relationships. Overreliance on *P* values and weak or misaligned measures further compound these risks. We urge researchers to be explicit about design logic, transparent about bias, and cautious when translating statistical associations into conclusions.

DISCLAIMER

The views expressed in this manuscript are solely those of the authors and do not necessarily reflect those of the Uniformed Services University of the Health Sciences, the Henry M. Jackson Foundation, or the United States Department of War.

REFERENCES

1. Peck EC, Lebensohn-Chialvo F, Fogarty CT. teaching family-oriented care to family medicine residents: evaluation of a family skills curriculum. *Fam Syst Health*. 2022;40(1):87–92. doi:10.1037/fsh0000659
2. Marbeen M, Freeman TR, Terry AL. Focused practice in family medicine: quantitative study. *Can Fam Physician*. 2022;68(12):905–914. doi:10.46747/cfp.6812905
3. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: a brief

primer. *BMJ Evid Based Med*. 2018;23(1):17–19. doi:10.1136/ebmed-2017-110800
4. Dyrbye LN, Szydlo DW, Downing SM, Sloan JA, Shanafelt TD. Development and preliminary psychometric properties of a well-being index for medical students. *BMC Med Educ*. 2010;10(1). doi:10.1186/1472-6920-10-8
5. Panagioti M, Panagopoulou E, Bower P, et al. Controlled interventions to reduce burnout in physicians: a systematic review and meta-analysis. *JAMA Intern Med*. 2017;177(2):195–205. doi:10.1001/jamainternmed.2016.7674
6. Dannels SA. Research Design. In: Hancock GR, Stapleton LM, Mueller RO, eds. *The Reviewer's Guide to Quantitative Methods in the Social Sciences*. Routledge; 2019:402–416
7. Omair A. Selecting the appropriate study design for your research: descriptive study designs. *J Health Spec*. 2015;3(3):153. doi:10.4103/1658-600X.159892
8. Price PC, Jhangiani RS, Chiang I-C. *Research Methods in Psychology*. BCcampus; 2014.
9. Onwuegbuzie AJ, McLean JE. Expanding the framework of internal and external validity in quantitative research. *Res Sch*. 2003;10(1):71–89.
10. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin; 2002.
11. Asan O, Montague E. Using video-based observation research methods in primary care health encounters to evaluate complex interactions. *Inform Prim Care*. 2014;21(4):161–170. doi:10.14236/jhi.v21i4.72
12. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23–34. doi:10.20982/tqmp.08.1.p023
13. Artino AR Jr, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. *Med Teach*. 2014;36(6):463–474. doi:10.3109/0142159X.2014.889814
14. Kotcher J, Maibach E, Miller J, et al. Views of health professionals on climate change and health: a multinational survey study. *Lancet Planet Health*. 2021;5(5):e316–e323. doi:10.1016/S2542-5196(21)00053-X
15. Bain L, Kennedy C, Archibald D, LePage J, Thorne C. A training program designed to improve interprofessional knowledge, skills and attitudes in chronic disease settings. *J Interprof Care*. 2014;28(5):419–425. doi:10.3109/13561820.2014.898622
16. Privitera GJ. *Essential Statistics for the Behavioral Sciences*. Sage; 2017.
17. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE guide No. 59. *Med Teach*. 2011;33(12):961–973. doi:10.3109/0142159X.2011.595435
18. Leppink J, van den Heuvel A. The evolution of cognitive load theory and its application to medical education. *Perspect Med Educ*. 2015;4(3):119–127. doi:10.1007/s40037-015-0192-x
19. Sarraf-Yazdi S, Pisupati A, Goh CK, et al. A scoping review and theory-informed conceptual model of professional identity formation in medical education. *Med Educ*. 2024;58(10):1151–1165. doi:10.1111/medu.15399
20. Syed M, McLean KC. Understanding identity integration: theoretical, methodological, and applied issues. *J Adolesc*. 2016;47(1):109–118. doi:10.1016/j.adolescence.2015.09.005
21. Koponen J, Pyörälä E, Isotalus P. Comparing three experiential learning methods and their effect on medical students' attitudes to learning communication skills. *Med Teach*. 2012;34(3):e198–207. doi:10.3109/0142159X.2012.642828
22. Gutersohn C, Schweingruber S, Haudenschild M, Huber M, Greif R, Fuchs A. Self-directed learning versus traditional instructor-led learning for education on a new anaesthesia workstation: a noninferiority, randomised, controlled trial. *Br J Anaesth*. 2025;135(4):990–996. doi:10.1016/j.bja.2025.03.043
23. Johnson AK, Blackstone SR, Simmons W, Skelly A. Assessing burnout and interest in wellness programs in physician assistant students. *J Physician Assist Educ*. 2020;31(2):56–62. doi:10.1097/JPA.0000000000000303
24. Ma TL, Bell K, Dong T, Durning SJ, Soh M. Military medical students' coping with stress to maintain well-being. *Mil Med*. 2023;188(Suppl 2):26–34. doi:10.1093/milmed/usac292
25. Yu B, Li Y, Wang J. Detecting Causal Language Use in Science Findings. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019; Stroudsburg, PA, USA, 4. https://www.aclweb.org/anthology/D19-1
26. Haber NA, Wieten SE, Rohrer JM, et al. Causal and associational language in observational health research: a systematic evaluation. *Am J Epidemiol*. 2022;191(12):2084–2097. doi:10.1093/aje/kwac137
27. Nakagawa S. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behav Ecol*. 2004;15(6):1044–1045. doi:10.1093/beheco/arh107
28. Abdi H. Holm's sequential Bonferroni procedure. In: Salkind NJ, ed. *Encyclopedia of Research Design*. Sage; 2010
29. Cabin RJ, Mitchell RJ. To Bonferroni or not to Bonferroni: when and how are the questions. *Bull Ecol Soc Am*. 2000;81(3):246–248.
30. Napierala MA. What is the Bonferroni correction? *AAOS Now*. 2012;6(4):40–41.
31. VanderWeele TJ, Mathur MB. Some desirable properties of the bonferroni correction: is the bonferroni correction really so bad? *Am J Epidemiol*. 2019;188(3):617–618. doi:10.1093/aje/kwy250
32. Oldfield M. *Unequal sample sizes and the use of larger control groups pertaining to power of a study*. Ministry of Defence UK Paper; 2016.
33. Kaplan D, George R. A study of the power associated with testing factor mean differences under violations of factorial invariance. *Struct Equ Modeling.*. 1995;2(2):101–118. doi:10.1080/10705519509539999
34. Lanz L, Thielmann I, Gerpott FH. Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *J Pers*. 2022;90(2):203–221. doi:10.1111/jopy.12662
35. Danner D, Aichholzer J, Rammstedt B. Acquiescence in personality questionnaires: Relevance, domain specificity,

and stability. *Journal of Research in Personality*. 2015;57:119–130. doi:10.1016/j.jrp.2015.05.004
36. Haghighat R. The development of the brief social desirability scale (BSDS). *EJOP*. 2007;3(4):417. doi:10.5964/ejop.v3i4.417
37. Greenwald HJ, Satow Y. A short social desirability scale. *Psychol Rep*. 1970;27(1):131–135. doi:10.2466/pr0.1970.27.1.131
38. Baines R, Regan de Bere S, Stevens S, et al. The impact of patient feedback on the medical performance of qualified doctors: a systematic review. *BMC Med Educ*. 2018;18(1). doi:10.1186/s12909-018-1277-0
39. McGlynn EA. Choosing and evaluating clinical performance measures. *Jt Comm J Qual Improv*. 1998;24(9):470–479. doi:10.1016/s1070-3241(16)30396-0
40. Campbell SM, Braspenning J, Hutchinson A, Marshall M. Research methods used in developing and applying quality indicators in primary care. *Qual Saf Health Care*. 2002;11(4):358–364. doi:10.1136/qhc.11.4.358
41. Goldstein R, Zuckerman B. A perspective on 360-degree evaluations. *J Pediatr*. 2010;156(1):1–2. doi:10.1016/j.jpeds.2009.09.027
42. Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. *Am J Phys Med Rehabil*. 2007;86(10):845–852. doi:10.1097/PHM.0b013e318151ff5a
43. Toegel G, Conger JA. 360-Degree assessment: time for reinvention. *AMLE*. 2003;2(3):297–311. doi:10.5465/amle.2003.10932156
44. Schleicher DJ, Baumann HM, Sullivan DW, Yim J. Evaluating the effectiveness of performance management: a 30-year integrative conceptual review. *J Appl Psychol*. 2019;104(7):851–887. doi:10.1037/apl0000368
45. Kendrick DE, Clark MJ, Fischer I, Bohnen JD, Kim GJ, George BC. The reliability of resident self-evaluation of operative performance. *Am J Surg*. 2021;222(2):341–346. doi:10.1016/j.amjsurg.2020.11.054
46. Juvonen J, Nishina A, Graham S. Self-views versus peer perceptions of victim status among early adolescents. In: Juvonen J, Graham G, eds. *Peer Harassment in School: The Plight of the Vulnerable and Victimized*. Guilford Press; 2001:105–124